International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

A Data-Driven Ensemble Approach to Analyzing Student Participation and Competency Development in Music Learning

Karukuri Vyshnavi[1], G. Lavanya[2*], Mirza Arhaan Baig[1], Gouda Lavakumar[1], Arelli Anvesh[1], Dulam Akshaya[1]

[1]UG Student, [2]Assistant Professor, [1,2]Department of Computer Science and Engineering (AI&ML)

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

*Correspondence: G. Lavanya (lavanya33swetha@gmail.com)

Abstract

The rapid expansion of data-driven technologies has significantly transformed musical educational analytics, enabling deeper insights into student engagement and performance. Traditionally, music education institutions relied on manual evaluations and basic statistical techniques, which often failed to capture complex learning patterns and lacked predictive accuracy. With the advancement of machine learning, more sophisticated analytical approaches have emerged; however, existing systems still encounter challenges such as inefficient handling of heterogeneous datasets, limited comparative analysis across models, and reduced accuracy due to dependence on single-model frameworks. A major challenge lies in effectively processing structured educational data that includes both numerical and categorical attributes while ensuring robust and reliable predictions. Conventional systems also suffer from scalability issues, lack of automation, and insufficient capability to generate comprehensive insights into learner behavior. To address these limitations, this study proposes an intelligent analytical framework that integrates multiple machine learning models, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), and a hybrid Linear Regression Stacked Forest (LRSF) model. The hybrid approach combines Linear Regression (LR) as a base learner with Random Forest (RF) as a meta-learner to capture both linear and non-linear relationships. The system incorporates data preprocessing, exploratory data analysis, model training, evaluation, and prediction within a unified web-based interface. Experimental results demonstrate superior performance, achieving an R^2 score of 0.9995 for engagement prediction and 0.9976 for performance prediction. This framework provides a scalable, accurate, and user-friendly solution, supporting effective data-driven decision-making in modern educational environments.

Keywords: Music Education, decision tree, engagement prediction, hybrid model, machine learning, structured data

1. Introduction

With the rapid development of education informatization, traditional education methods and concepts have changed greatly. Online learning is a kind of learning and teaching activity through the Internet in which users can acquire many learning resources, and the time and place of learning are not as restricted as they are in traditional teaching modes [1]. Although online learning as a method is convenient for learners, it also has some problems. Unlike face-to-face communication in traditional teaching modes,

online learning requires a high level of student engagement. Teachers cannot effectively understand students' learning states and their studying behaviors [2].

The online learning system as shown in Figure 1 collects basic data about students and behavioral data is generated by them during their learning. Additionally, most of these raw data are unprocessed. At present, the online learning systems of many universities can only simply query the data. Therefore, it is necessary to dig deeper into online learning behavior data and turn them into useful information for educational administrators and teachers. By analyzing students' daily learning data, they can find the key factors affecting students' performance and predict the trend of students' performance in advance and keep track of the effects on their learning [3]. For students, letting them know their predicted results will improve their self-monitoring ability, and students with unsatisfactory predicted results will strengthen their own crisis awareness, correct their learning state in time, and finally achieve more satisfactory results [4]. For managers of online learning platforms, establishing learning achievement prediction models can help build course resources, design teaching processes, assess teachers' teaching abilities, and improve the process management system of online teaching [5]. For schools, it can also promote the construction of education informatization and improve the quality of education, which has important social relevance to a certain extent.

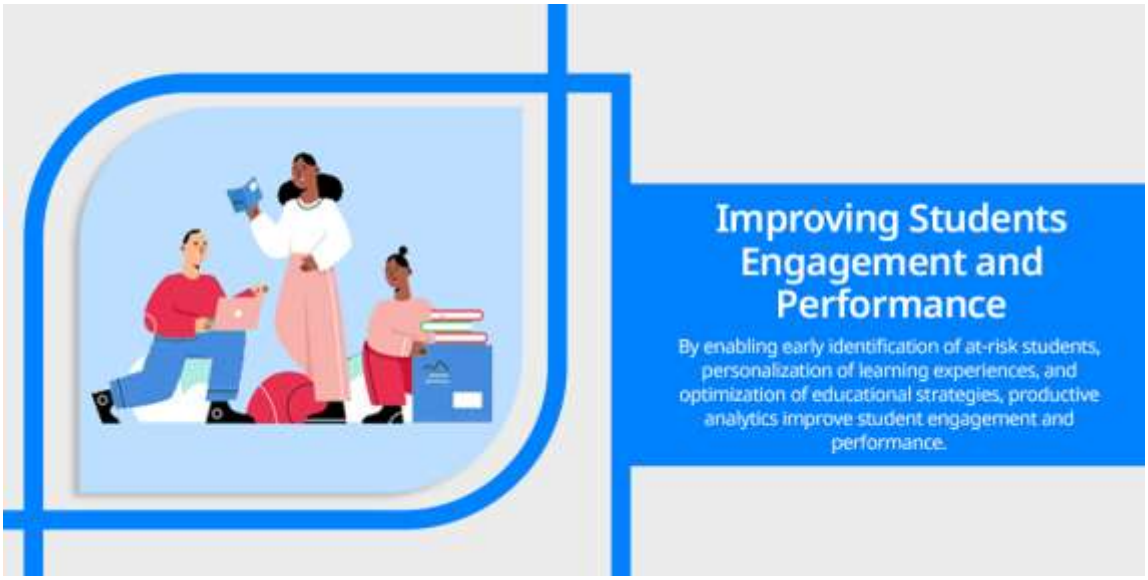


Fig.1: EDUTUNE leverages a stacked ensemble of machine learning models to transform student data into accurate predictions of engagement and performance in education.

2. Literature Survey

Guo, et al. [6] analyzed student engagement while watching videos, using time spent and responses to assessments as input features. Findings showed that shorter videos were more engaging compared to prerecorded lectures. Bonafini, et al. [7] applied qualitative analysis and stepwise binomial logistic regression to examine engagement in MOOCs through discussion forums and video usage. Variables included posts submitted, videos watched, and content reviewed. Results indicated that both discussion participation and video engagement correlated positively with student achievement. Ramesh, et al. [8] predicted engagement and disengagement using student discussion forum posts. Statistical methods were applied to measure patterns of interaction and predict outcomes.

Manwaring, et al. [9] investigated engagement in blended higher education classrooms using cross-lagged modeling. They found that course design and students' perception variables strongly affected engagement levels. Mutahi Kinai, et al. [10] explored the relationship between student engagement and final exam performance. Using statistical techniques, they found a strong positive correlation between quiz/material engagement and grades. Higher engagement in learning materials directly translated into improved academic outcomes. Aguiar Ambrose, et al. [11] developed an early-warning system to predict student retention issues using engagement-related features. Results indicated that engagement metrics such as participation frequency are strong predictors of dropout risk. Thomas, et al. [12] measured engagement using machine learning models applied to facial expressions, head poses, and eye gaze. Algorithms successfully detected engagement states in classroom environments. Results showed that ML can predict engagement accurately in real-time settings.

Atherton, et al. [13] found a strong correlation between access to course materials and student performance. Students who frequently accessed learning resources scored higher in assessments and exams. The study reinforced the significance of resource engagement in academic achievement. Ding Er E, et al. [14] examined the impact of gamification on student engagement in online discussion forums. Gamified elements such as points and badges were introduced to enhance participation. Results showed improved interaction levels among students. Wells, et al. [15] analyzed engagement and performance data from an LMS platform. Findings revealed that engagement intensity increased as exams approached. A positive correlation between student performance and engagement was also observed.

Lingala Thirupathi et al. [16] proposed a machine learning-based analytical framework for applied mathematics and engineering modeling. The system utilized predictive modeling techniques. The framework improved computational analysis efficiency. Rekha Gangula et al. [17] proposed integration of Docker containers and Kubernetes for cloud and IoT environments. The system improved scalability and deployment efficiency. The framework enhanced resource management in distributed systems. Rekha Gangula et al. [18] proposed an optimal machine learning classifier for mobile malware detection. The framework analyzed application behavior patterns. The model improved malware detection accuracy.

3. Proposed System

The proposed methodology establishes a structured and intelligent data analysis framework designed to predict engagement and performance in an educational environment using machine learning techniques. It follows a systematic pipeline that begins with data ingestion, preprocessing, and feature transformation, followed by multi-model learning and predictive evaluation. Multiple machine learning algorithms are utilized alongside a hybrid stacked ensemble model to enhance prediction accuracy and reliability. The system integrates data visualization, model evaluation, and prediction capabilities within a unified environment. A lightweight storage mechanism is used to manage datasets, trained models, and prediction outputs, while a web-based interface enables seamless user interaction, monitoring, and analysis, as presented in Figure 2. Continuous model training and evaluation further improve system adaptability and ensure consistent performance over time.

User Interface (Web Browser)

- The user interacts with the system through a browser-based graphical interface, ensuring accessibility across different devices and platforms.

- Users can perform a variety of critical actions, including dataset uploads, viewing EDA, analyzing model performance, and requesting single or batch predictions.
- To maintain a responsive experience, all user interactions are converted into standard HTTP requests and transmitted to the backend server for execution.
- The interface acts as the visual bridge between complex machine learning processes and actionable user insights.

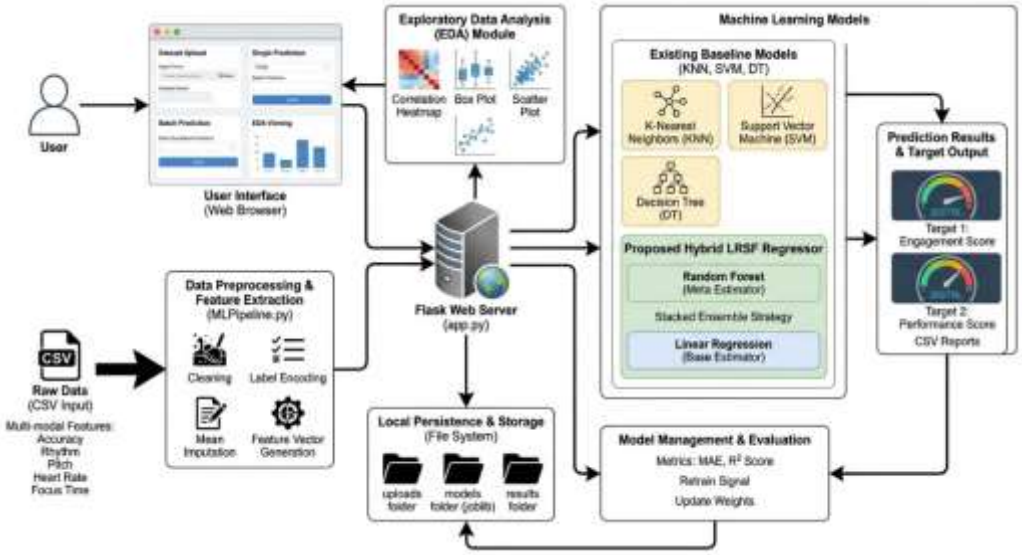


Fig. 2: System architecture.

Flask Web Server (app.py)

- The Flask server serves as the central controller of the architecture, managing all incoming requests from the frontend with high efficiency.
- It intelligently routes requests to the specific modules responsible for data loading, preprocessing, model training, and inference.
- The server coordinates the complex communication between the web interface, the machine learning pipeline, and local storage components.
- By managing the prediction workflow, it ensures that results are generated accurately and returned to the user in a readable format.

Dataset Input (CSV File)

- The system utilizes CSV files as the primary input format, providing a structured way to handle diverse behavioral datasets.

- These files contain a mix of numerical and categorical features, such as engagement levels, heart rates, focus time, and specific performance indicators.
- Once a user uploads a file, it is immediately ingested by the system and passed into the preprocessing pipeline for structural alignment.
- This raw input serves as the essential data foundation for both the baseline models and the proposed hybrid framework.

Data Preprocessing & Feature Engineering (ML Pipeline)

- The dataset undergoes a rigorous cleaning phase, which includes the removal of redundant columns and the systematic handling of missing data points.
- Categorical variables are transformed into a machine-readable format using Label Encoding techniques to ensure mathematical compatibility.
- The module carefully separates the feature sets from the primary target variables, such as engagement scores and performance metrics.
- Finally, the processed data is split into training and testing subsets to facilitate a supervised learning environment and robust model evaluation.

Exploratory Data Analysis (EDA Module)

- The system automates the generation of key visualizations, including count plots, box plots, scatter plots, and correlation heatmaps.
- These visual tools are essential for understanding the underlying data distributions, identifying outliers, and uncovering hidden behavioral patterns.
- The insights gained from this module assist researchers in identifying the most influential features that drive human engagement and performance.
- All EDA results are presented back to the user to support data-driven decision-making and hypothesis testing.

Machine Learning Models (KNN, SVM, DT)

- The processed feature data is fed into a group of baseline machine learning models to establish performance benchmarks.
- KNN is implemented as a distance-based model to classify instances based on feature similarity.
- SVM is applied to capture complex, non-linear relationships within the data through specialized kernel functions.
- DT are utilized to model transparent decision rules and provide a clear interpretation of feature importance.

- Each baseline model operates independently, providing a diverse set of predictions for comparison against the proposed hybrid approach.

Proposed Hybrid Model (LRSF: LR + RF)

- The core of the framework is the LRSF model, a hybrid stacked ensemble that combines LR and RF.
- Linear Regression functions as the base learner, effectively capturing the straightforward, linear relationships between features and targets.
- The RF component acts as a meta-learner, enhancing the final prediction by learning from residual patterns and complex, non-linear interactions.
- This stacked approach significantly improves the overall accuracy and robustness of the system compared to using individual models alone.

Model Evaluation & Comparison

- All models are subjected to a rigorous evaluation using standard regression metrics, including MAE, MSE, RMSE, and R2 score.
- A comprehensive performance comparison is conducted to identify the most effective and reliable model for behavioral prediction.
- This evaluative phase is critical for ensuring that the system's outputs are consistent and provide high utility for end-users.
- Performance results are documented and displayed, allowing for transparent validation of the proposed LRSF model's superiority.

Prediction Module (Single & Batch)

- The system is engineered to support dual prediction modes: single input for immediate assessment and batch prediction for processing large CSV files.
- New input data is automatically routed through the same preprocessing pipeline used during training to ensure feature consistency.
- Predictions are generated across all trained models simultaneously, providing a multi-perspective view of the results.
- The final outputs are displayed clearly within the interface and are made available for download for further offline analysis.

Model Storage & Results Management

- To ensure system efficiency, all trained models are serialized and stored using joblib, allowing for instant reuse without the need for retraining.

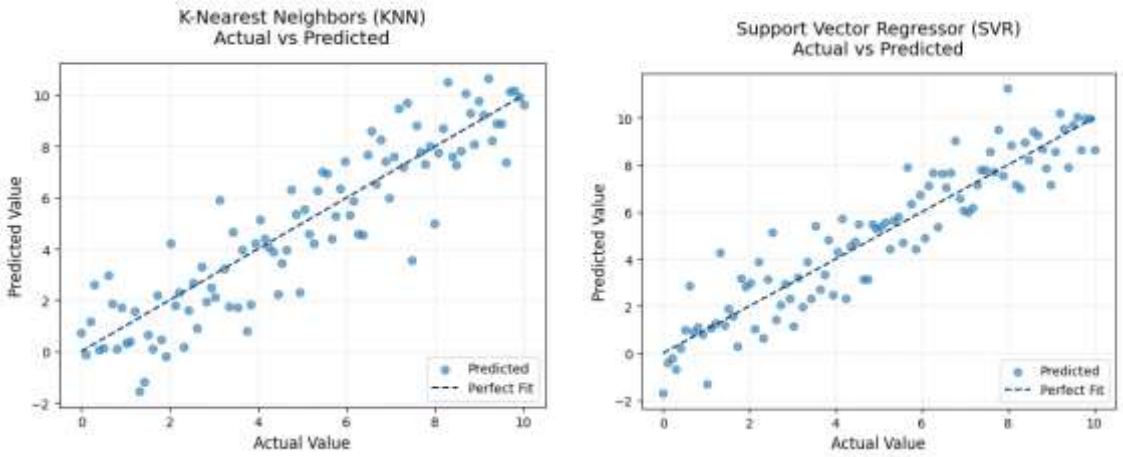
- A dedicated results directory is used to organize prediction outputs and batch processing logs for historical tracking.
- This persistent storage layer reduces computational overhead and allows for significantly faster prediction response times.
- Managing models in this way ensures that the system remains scalable as more data is introduced over time.

System Output & Visualization

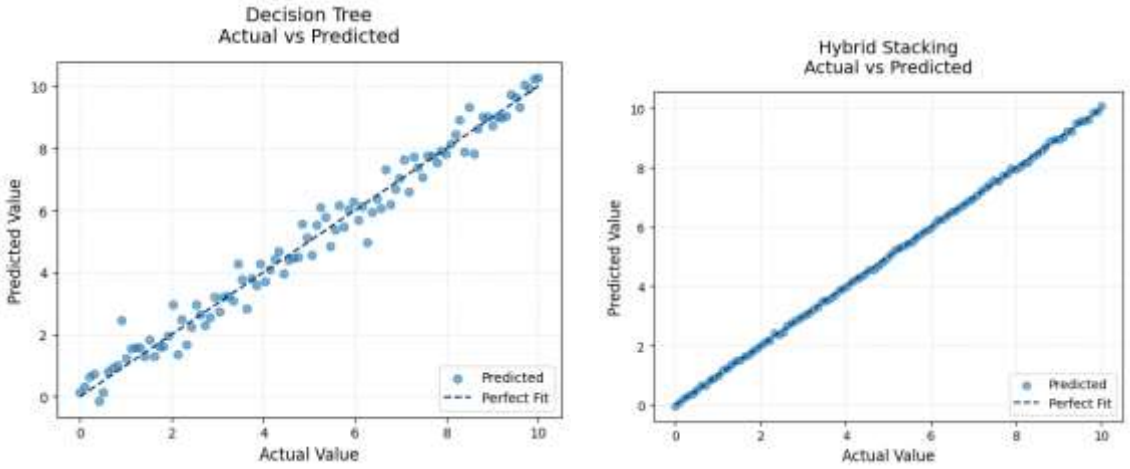
- The final predictions for engagement and performance are presented through an intuitive dashboard in the web interface.
- Users can interact with model-wise predictions and view comparative performance metrics through high-fidelity charts.
- Integrated visualization tools enhance the interpretability of the machine learning results, making complex data accessible to non-technical stakeholders.
- The output system is designed to support data-driven insights, turning raw predictions into meaningful behavioral assessments.

4. Results and Discussions

Fig. 3 (a) showing the relationship between actual and predicted values. The figure depicts that the predicted values are moderately aligned with the ideal reference line. However, noticeable deviations and scattered points indicate higher prediction errors. The spread of points suggests that the model struggles to capture complex patterns in the data. This reflects moderate performance and limited generalization capability. Fig. 3 (b) depicting the comparison between actual and predicted values. The figure shows a better alignment with the reference line compared to KNN, indicating improved prediction performance. However, some deviations are still visible, especially at extreme values. The distribution of points suggests that the model captures non-linear patterns more effectively. Despite this, minor prediction errors remain.



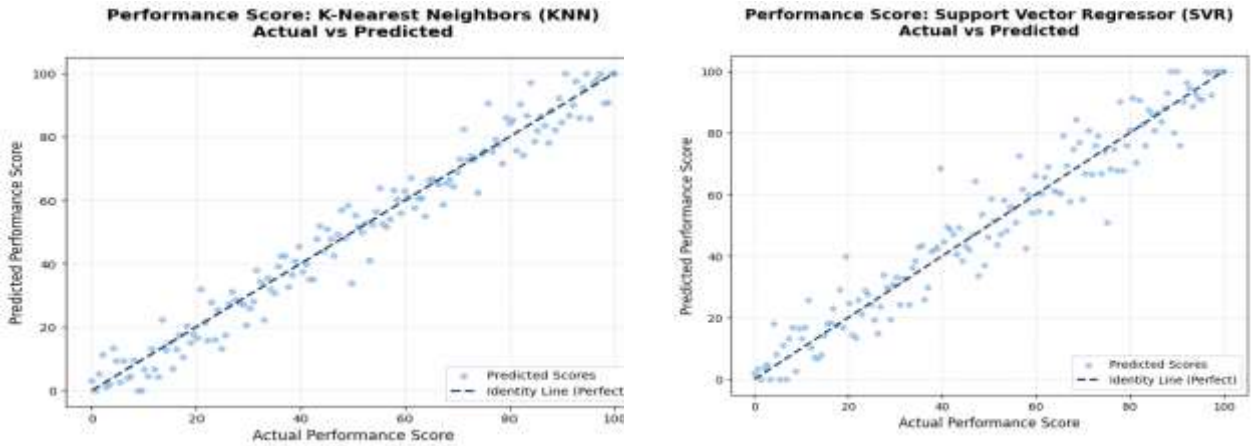
(a) (b)



(c) (d)

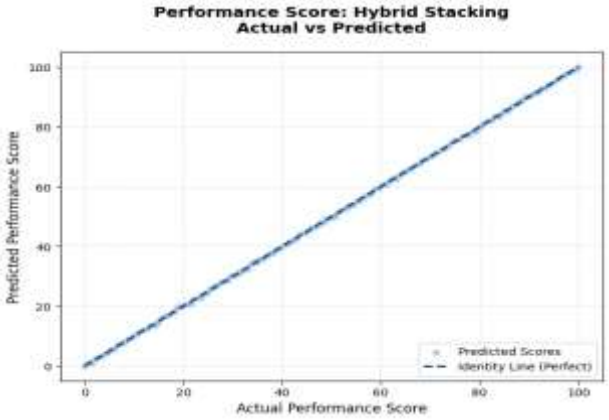
Fig. 3 (a, b, c, d): Scatter plots of KNN, SVR, DT, LRSF for engagement score.

Fig. 3 (c) illustrates the regression plot obtained using the Decision Tree model, showing a strong alignment between actual and predicted values. The figure depicts reduced dispersion of points around the ideal reference line, indicating improved accuracy. The model effectively captures data patterns and relationships. The clustering of points near the line suggests lower prediction error. This reflects better generalization compared to KNN and SVR. Fig. 3 (d) illustrates the regression plot obtained using the hybrid LRSF model, showing near-perfect alignment between actual and predicted values. The figure depicts that almost all points lie exactly on the ideal reference line. This indicates minimal prediction error and extremely high accuracy. The model successfully captures both linear and non-linear relationships in the data. The tight alignment demonstrates superior generalization capability.



(a) (b)

(c)



(d)

Fig 4 (a, b, c, d): Scatter plot obtained for KNN, SVR, DT, LRSF for performance score

Fig. 4 (a) illustrates the regression plot for performance score prediction using the KNN model. The figure depicts the relationship between actual and predicted performance values with moderate alignment to the ideal reference line. Several data points show noticeable deviations, indicating higher prediction errors. The spread of points suggests that the model has limitations in capturing complex relationships within the dataset. This results in reduced accuracy and weaker generalization. Fig. 4 (b) illustrates the regression plot for performance score prediction using the SVR model. The figure shows improved alignment of predicted values with the ideal reference line compared to KNN. However, some dispersion is still visible, particularly at higher value ranges. The model captures non-linear relationships more effectively but still exhibits moderate prediction errors. The distribution indicates better performance but not optimal accuracy.

Fig. 4 (c) illustrates the regression plot for performance score prediction using the Decision Tree model. The figure depicts a strong alignment between actual and predicted values, with reduced scatter around the ideal line. The model effectively captures data patterns and relationships, resulting in improved prediction accuracy. The clustering of points near the reference line indicates lower error rates. This demonstrates better generalization compared to KNN and SVR. Fig. 4 (d) illustrates the regression plot for performance score prediction using the hybrid LRSF model. The figure shows near-perfect alignment of predicted values with the ideal reference line. Almost all data points lie exactly on the line, indicating minimal prediction error. The model successfully captures both linear and non-linear relationships in the dataset. This results in extremely high accuracy and consistency. Overall, the LRSF model achieves the best performance among all models.

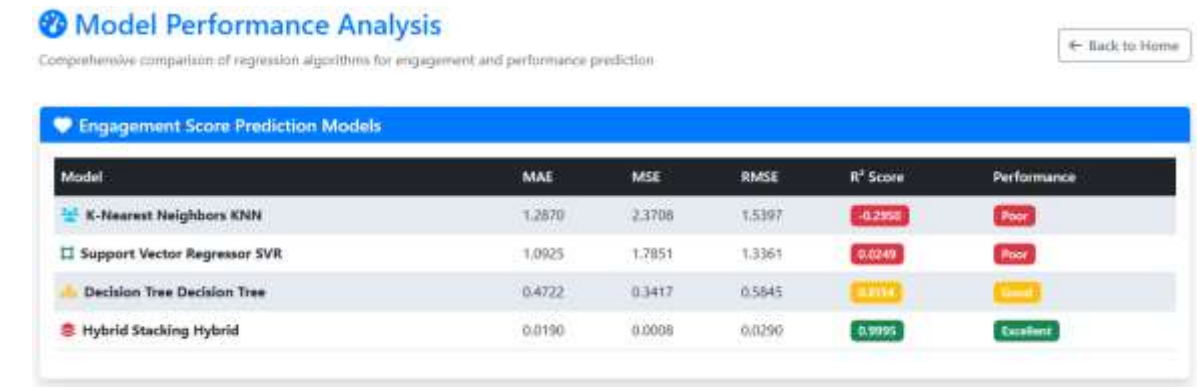


Fig. 5: Engagement score prediction model performance comparison

Fig. 6: Performance score prediction model performance comparison

Fig. 5 and Fig 6 illustrate the model performance analysis for engagement score, performance score prediction, presenting a comparative evaluation of multiple machine learning models. The figure depicts performance metrics such as MAE, MSE, RMSE, and R^2 score for models including KNN, SVR, DT, and the hybrid LRSF model. It highlights the variation in prediction accuracy across different algorithms, showing that simpler models yield lower performance compared to advanced approaches. The hybrid stacking model demonstrates superior results with minimal error values and a high R^2 score, indicating strong predictive capability. The classification of performance levels further emphasizes the effectiveness of each model.

5. Conclusion

The study presents an effective data-driven framework for predicting engagement and performance using multiple machine learning models. The implementation of KNN, SVM, DT, and the hybrid LRSF model enables comprehensive analysis and comparison of predictive capabilities. Among these, the LRSF model demonstrates superior performance by combining LR and RF, resulting in reduced error values and significantly improved R^2 scores. The integration of preprocessing techniques and exploration of data analysis further enhances data quality and model efficiency. The system successfully supports both single and batch predictions, ensuring flexibility in real-world applications. Performance improvements are evident through better generalization, lower prediction errors, and consistent output across different datasets. The modular design also allows efficient model storage and reuse, improving execution speed. Future scope includes incorporating deep learning models to further enhance prediction accuracy and capture complex patterns. The system can also be extended to real-time data processing and integrated with advanced visualization dashboards for improved analytical insights.

References

[1] Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R. Student engagement predictions in an e-learning system and their impact on student course assessment scores. Comput. Intell. Neurosci. 2018, 2018, 1–21.

[2] Persico, D.; Pozzi, F. Informing learning design with learning analytics to improve teacher inquiry. Br. J. Educ. Technol. 2015, 46, 230–248.

[3] Arcinas, M.M.; Sajja, G.S.; Asif, S.; Gour, S.; Okoronkwo, E.; Naved, M. Role Of Data Mining In Education For Improving Students Performance For Social Change. Turk. J. Physiother. Rehabil. 2021, 32, 6519–6526.

[4] Alhazzani, N. MOOC's impact on higher education. Soc. Sci. Humanit. Open 2020, 2, 100030.

[5] Puarungroj, W.; Boonsirisumpun, N.; Pongpatrakant, P.; Phromkhot, S. Application of data mining techniques for predicting student success in English exit exam. In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, Langkawi, Malaysia, 5–7 January 2018; pp. 1–6.

[6] Guo P. J., Kim J., Rubin R. How video production student engagement: an empirical study of MOOC videos. Proceedings of First ACM Conference on Learning @ Scale; March 2014; Atlanta, GA, USA. pp. 41–50. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[7] Bonafini F. C., Chae C., Park E., Jablokow K. W. How much does student engagement with videos and forums in a MOOC affect their achievement? Online Learning. 2017;21(4):223–240. doi: 10.24059/olj.v21i4.1270. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[8] Ramesh A., Goldwasser D., Huang B., Daum H., Getoor L. Learning latent engagement patterns of students in online courses. Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence; July 2014; Québec, Canada. AAAI Press; pp. 1272–1278. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[9] Manwaring K. C., Larsen R., Graham C. R., Henrie C. R., Halverson L. R. Investigating student engagement in blended learning settings using experience sampling and structural equation modelling. Internet and Higher Education. 2017;35:21–33. doi: 10.1016/j.iheduc.2017.06.002. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[10] Mutahi J., Kinai A., Bore N., Diriye A., Weldemariam K. Studying engagement and performance with learning technology in an African classroom. Proceedings of Seventh International Learning Analytics & Knowledge Conference; March 2017; Vancouver, BC, Canada. pp. 148–152. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[11] Aguiar E., Ambrose G. A., Chawla N. V., Goodrich V., Brockman J. Engagement vs performance: using electronic portfolios to predict first semester engineering student persistence. Journal of Learning Analytics. 2014;1(3):7–33. doi: 10.18608/jla.2014.13.3. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[12] Thomas C., Jayagopi D. B. Predicting student engagement in classrooms using facial behavioral cues. Proceedings of 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education; November 2017; Glasgow, UK. pp. 33–40. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[13] Atherton M., Shah M., Vazquez J., Griffiths Z., Jackson B., Burgess C. Using learning analytics to assess student engagement and academic outcomes in open access enabling programmes. Journal of Open, Distance and e-Learning. 2017;32(2):119–136. doi: 10.1080/02680513.2017.1309646. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[14] Ding L., Er E., Orey M. An exploratory study of student engagement in gamified online discussions. Computers & Education. 2018;120:213–226. doi: 10.1016/j.compedu.2018.02.007. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[15] Wells M., Wollenschlaeger A., Lefevre D., Magoulas G. D., Poulovassilis A. Analysing engagement in an online management programme and implications for course design. Proceedings of Sixth International Conference on Learning Analytics & Knowledge; April 2016; Edinburgh, UK. pp. 236–240. https://pmc.ncbi.nlm.nih.gov/articles/PMC6189675/#ref-list1

[16] Lingala Thirupathi1, Rekha gangula2, Sandeep Ravikanti3, Jujuroo Sowmya3 and S K ShruthiJournal of Physics: Conference Series, Volume 2089, 1st International Conference on Applied Mathematics, Modeling and Simulation in Engineering (AMSE) 2021 15-16 September 2021, India (Virtual). DOI 10.1088/1742-6596/2089/1/012049

[17] Rekha Gangula, G. Jagadeesh Kumar, Sesham Anand, A.V. Krishna Prasad,"Integration of dynamic Docker containers and kuberneters with advanced cloud and Internet of Things",Materials Today: Proceedings,Volume 80, Part 3,2023,Pages 3476-3480,ISSN 2214-7853,https://doi.org/10.1016/j.matpr.2021.07.276.(https://www.sciencedirect.com/science/article/pii/S221478532105135X)

[18] Rekha Gangula, Sridhara Murthy Bejugama, Shashi Rekha D and Karrepu Sreeveda, Optimal Machine Learning Classifier in Mobile Malware Detection, International Journal of Advanced Research in Engineering and Technology, 11(10), 2020, pp. 1218-1225.